



Comparative assessment of multiple criteria for the *in silico* prediction of cross-reactivity of proteins to known allergens



Henry P. Mirsky*, Robert F. Cressman Jr.¹, Gregory S. Ladics¹

DuPont Pioneer, Route 141, Henry Clay Bldg, #400, Wilmington, DE 19880-0400, USA

ARTICLE INFO

Article history:

Received 4 February 2013

Available online 9 August 2013

Keywords:

Allergenic cross-reactivity

In silico

Bioinformatics

Assessment criteria

False positives

False negatives

Sensitivity

Specificity

ABSTRACT

Genetically modified crops are becoming important components of a sustainable food supply and must be brought to market efficiently while also safeguarding the public from cross-reactivity of novel proteins to known allergens. Bioinformatic assessments can help to identify proteins warranting further experimental checks for cross-reactivity. This study is a large-scale *in silico* evaluation of assessment criteria, including searches for: alignments between a query and an allergen having $\geq 35\%$ identity over a length ≥ 80 ; any sequence (of some minimum length) found in both a query and an allergen; any alignment between a query and an allergen with an *E*-value below some threshold. The criteria and an allergen database (AllergenOnline) are used to assess 27,243 *Viridiplantae* proteins for potential allergenicity. (A protein is classed as a “real allergen” if it exceeds a test-specific level of identity to an AllergenOnline entry; assessment of real allergens in the query set is against a reduced database from which the identifying allergen has been removed.) Each criterion’s ability to minimize false positives without increasing false negative levels of current methods is determined. At best, the data show a reduction in false positives to $\sim 6\%$ (from $\sim 10\%$ under current methods) without any increase in false negatives.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Prior to commercialization, genetically modified (GM) crops undergo an extensive safety evaluation. One component of this evaluation is the assessment of the newly expressed protein(s) for allergenic potential. Currently, no single factor is recognized as an identifier for protein allergenicity. However, *in silico* screening of genetic modifications has become mandatory and may help to identify novel proteins possessing possible allergenic cross-reactivity – and therefore warranting subsequent experimental tests for confirmation (e.g., IgE-binding studies, pepsin digestibility studies, etc.). These data (and their quality) are then weighed to reach a conclusion regarding risk (Codex Alimentarius Commission, 2009).

Near-identity of candidate protein sequences with known allergens will certainly identify them as such, if that is the case, and serum testing would presumably be unnecessary. For those query sequences that do not clearly match known allergens, bioinformatic assessment, although not a replacement for empirical assessment of IgE binding, can provide insight into possible cross-reactivity, as sequences sharing a high degree of identity

often share immunologically relevant topology (Aalberse, 2000; Aalberse et al., 2001; Goodman et al., 2008.) Traditionally, such analysis consists of two components: A search for short linear epitopes as well as comparison of primary amino acid sequences using FASTA (Pearson and Lipman, 1988) or BLAST (Altschul et al., 1997) to locate possible shared conformations (Cressman and Ladics, 2009). Recent publications (EFSA, 2011; Goodman, 2008; Goodman and Tetteh, 2011; Harper et al., 2012; Herman et al., 2009; Ladics et al., 2011; Young et al., 2012) have argued that the standard search for a sequence of eight or more amino acids found in both the query and a known allergen provides little value; however it has been demonstrated that insertion of a short stretch of amino acids derived from a known allergen into the correct conformational context can result in an increase in specific IgE binding (Klinglmayr et al., 2009). Additionally, the default local alignment search criteria have been constrained by the imposition of a defined threshold ($\geq 35\%$ sequence identity over an alignment length ≥ 80) (Ladics et al., 2007). This constraint neglects many of the features that help to define relevant homologies between sequences, features incorporated into the algorithms themselves (e.g., *E*-value [expectation value], which is a measure of the relatedness between protein sequences). A small *E*-value (e.g., 10^{-8}) indicates a potential biologically relevant similarity in the context of potential allergenic cross-reactivity; large *E*-values (>1.0) represent random alignments that do not possess biologically relevant similarity (Pearson, 2000; Silvanovich et al., 2009).

* Corresponding author. Fax: +1 302 695 3075.

E-mail addresses: henry.mirsky@pioneer.com (H.P. Mirsky), robert.f.cressman@pioneer.com (R.F. Cressman Jr.), gregory.s.ladics@usa.dupont.com (G.S. Ladics).

¹ Fax: +1 302 695 3075.

In order to investigate the refinement of current bioinformatic practices for assessing potential cross-reactivity, we have undertaken a systematic investigation of the effect of various *in silico* assessment criteria on the ability to distinguish real allergens from non-allergens. The various criteria include one or more of the following components: (1) A check for the presence of an alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 (Codex Alimentarius Commission, 2009); (2) a check for the presence of a sequence, at or above a specified length, found in both a query protein and an allergen; and (3) a check for the presence of an alignment between a query protein and an allergen with an *E*-value at or below some specified threshold. While there have been previous investigations into the way some of these criteria affect detection rates (Silvanovich et al., 2009), to our knowledge there have been no studies on this scale. Our analyses have uncovered numerous potential allergenic assessment criteria that are able to lower the current percent of false positives without raising the current percent of false negatives – at best, a reduction in the percent of false positives from $\sim 10\%$ to $\sim 6\%$ (a $>40\%$ reduction) is achieved without any increase in the percent of false negatives and with only minor changes to the existing criteria.

To date, there have been no known instances of transgenic proteins in GM crops inducing responses in humans. As food demand increases while environmental pressures grow, it is essential that the adoption of promising crop traits not be delayed by extraneous

experimental investigations caused by positive but misleading bioinformatic matches. Therefore, *in silico* predictions of possible cross-reactivity should be as useful as possible, and it is hoped that the results provided here will help to serve that end.

2. Methods and materials

2.1. Query proteins

The entire set (31,897 sequences) of reviewed *Viridiplantae* proteins was obtained from UniProtKB/Swiss-Prot (<http://www.uniprot.org>). Under current Codex guidelines, an alignment of a protein to an allergen must possess a length ≥ 80 ; therefore, sequences with lengths smaller than 80 amino acids were removed from further consideration. Additionally, any sequence annotated as a “fragment” was removed. The final query protein dataset contained 27,243 protein sequences. See Fig. 1.

2.2. Allergens

Version 12 (January 2012) of the AllergenOnline database (<http://www.allergenonline.org>), a peer-reviewed allergen database compiled by the Food Allergy Research and Resource Program (FARRP) (<http://farrp.unl.edu>) at the University of Nebraska, Lincoln, was used for analyses. AllergenOnline v12 contains 1603 protein sequences.

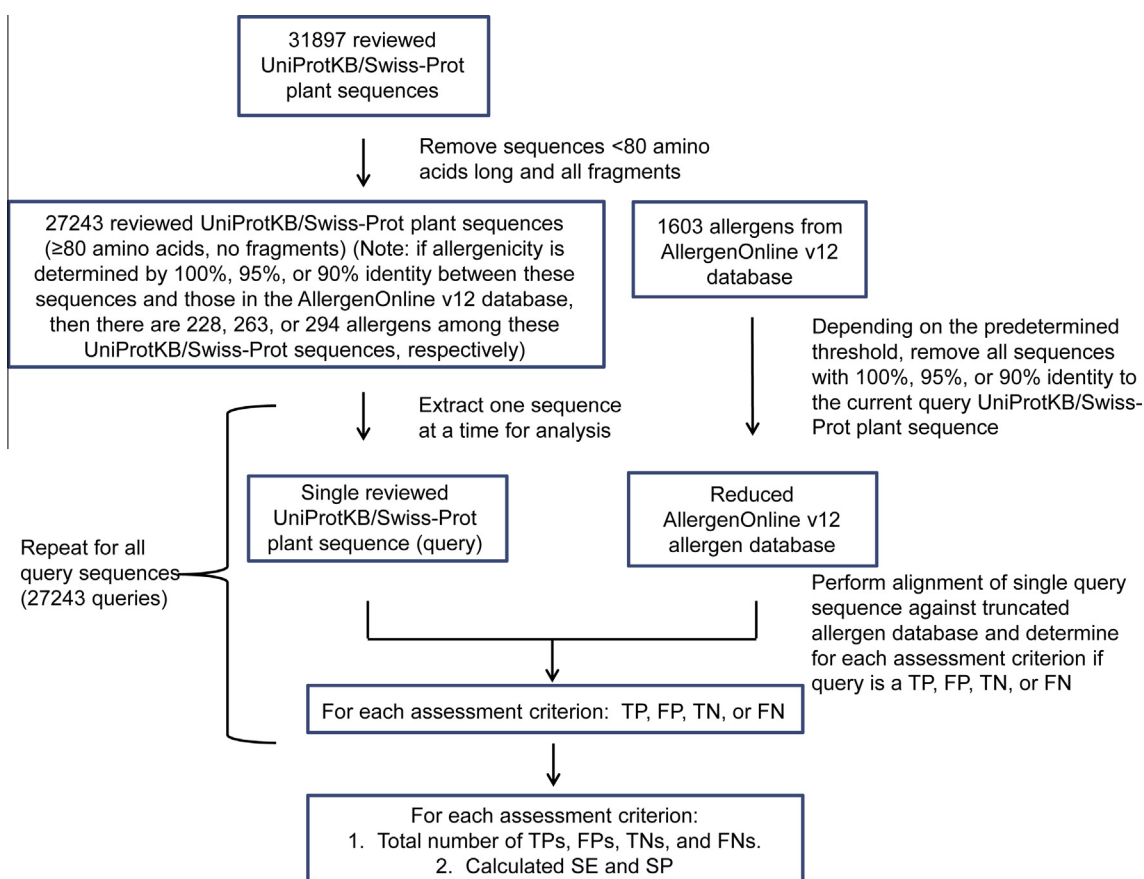


Fig. 1. Process flow chart. The entire set of reviewed *Viviplantae* proteins (31,897 proteins) was downloaded from UniProtKB/Swiss-Prot (<http://www.uniprot.org>). All proteins smaller than 80 amino acids in length, as well as all fragments, were removed. The first of the remaining 27,243 proteins was selected. This protein was compared against the 1603 proteins in the Food Allergy Research and Resource Program's (FARRP) AllergenOnline (<http://www.allergenonline.org>) (Version 12) (AllergenOnline v12) dataset and declared a “real allergen” if sequence identity was 100%, $\geq 95\%$, or $\geq 90\%$, depending on the test. The matching sequences in AllergenOnline v12, if any, were removed from AllergenOnline v12 to form a reduced allergen database. The query protein was evaluated for allergenicity against the reduced allergen database for each specified allergenic assessment criterion and declared either a true positive (TP), false positive (FP), true negative (TN), or false negative (FN). The original AllergenOnline v12 database was restored, and the process was repeated for each of the remaining query proteins. The total number of TPs, FPs, TNs, and FNs produced using each allergenic assessment criterion was used to calculate the criterion's sensitivity (SE) and specificity (SP).

2.3. Determination of “real” allergenicity using query set proteins

Prior to any analysis, each UniProtKB/Swiss-Prot query protein was assessed for its “real allergenicity” by comparison to AllergenOnline v12. For the first set of evaluations a query protein was considered a “real allergen” if it possessed 95% identity with a sequence in AllergenOnline v12; otherwise it was considered a non-allergen. Of the 27,243 sequences checked, 263 were thereby designated real allergens. Subsequently all evaluations were repeated using both a very stringent requirement of 100% identity followed by a less stringent requirement of 90% identity to determine real allergenicity, yielding 228 and 294 allergens respectively.

2.4. Creation of reduced allergen lists

Because, by definition, any protein in the query protein dataset that had been designated a real allergen is identical or nearly identical (90–100% identity) to a protein in AllergenOnline v12, it would of course always assess positively as a possible allergen regardless of the assessment criterion used. Since the goal of this study was the determination of criteria to assess unknown proteins for potential allergenicity, each real allergen in the query protein dataset was assessed for possible allergenicity by using a slightly reduced version of AllergenOnline v12 from which the sequences with 100%, 95%, or 90% identity (as appropriate for the specific test) had been removed. See Fig. 1.

2.5. Classification of assessed query proteins

Assessment for potential allergenicity of the proteins in the query protein dataset was made according to various criteria as shown in Tables 1 and 2. Any query protein that was assessed as potentially allergenic was declared a *true positive* (TP) if a real allergen and a *false positive* (FP) if not. Conversely, any query protein that was assessed as non-potentially-allergenic was declared a *false negative* (FN) if a real allergen and a *true negative* (TN) if not.

2.6. Metrics

The criteria for allergenic assessment classify a real allergen as either a TP or a FN. The ability to distinguish the two types is quantified via the *sensitivity* (SE) (Baldi et al., 2000), where

$$SE = TP / (TP + FN) \quad (1)$$

If SE = 0, this indicates that all allergens were incorrectly designated *false*, while if SE = 1, all allergens were correctly designated *true*.

Conversely, a non-allergen may be called either a TN or a FP; *Specificity* (SP) is given as (Baldi et al., 2000; Burset and Guigo, 1996).

$$SP = TN / (TN + FP). \quad (2)$$

SP also assumes a range from 0 to 1; if SP = 1, this means that there are no false positives.

2.7. Comparison of test criteria

The proteins in the query protein dataset at each level of real allergenicity determination (100%, 95%, and 90%) were assessed using various criteria, including two alternative criteria under the Codex guidelines (Established Method 1 and Established Method 2, Table 1) and numerous proposed criteria (Table 2), producing criterion-specific numbers of TPs, FNs, TNs, FPs, and consequently SEs and SPs. The SE and SP for Established Method 2 were then compared against the SE and SP for each proposed criterion. If the SE for the proposed criterion was lower than the SE for Established Method 2, then the proposed criterion was considered inferior to Established Method 2 for SE (and likewise for SP). Otherwise, the proposed criterion was considered comparable or superior for the relevant metric. The “best” proposed criteria were the ones that yielded the greatest improvement in SP without any reduction in SE.

2.8. Hardware and software

Production of reduced allergen lists, analyses per pre-specified criteria, and determination of TPs, FPs, TNs, FNs, SEs, and SPs were achieved with in-house Perl (ActiveState version 5.12.3, <http://www.activestate.com>) scripts operating on a personal computer. The FASTA alignment program (Pearson and Lipman, 1988) (version 35.4.4) was used for processing alignments of query proteins against the AllergenOnline v12 allergens.

3. Results

3.1. Determination of the real allergenicity of proteins in the query protein dataset

All of the results shown below are for the case where the actual allergenicity of a query protein was determined at the 95% level (i.e., a query protein was identified as a real allergen if it possesses 95% (or greater) identity to a protein in the AllergenOnline v12 database). This non-exactness was permitted to account for those instances where a few amino acid changes would otherwise result in a classification of a protein as a non-allergen.

3.2. Allergenic assessment using Established Method 1 produces mostly false positives

The Food and Agriculture Organization (FAO)/World Health Organization (WHO) Codex Principles and Guidelines on Foods Derived from Biotechnology (2001) originally recommended a positive assessment of possible allergenicity for a query protein if there is (1) any alignment between any 80-amino-acid-long window on the query protein and an allergen with $\geq 35\%$ sequence identity over an alignment length ≥ 80 or (2) any sequence of six or more amino

Table 1
The 2 “Established Methods” used to assess allergenicity, the criteria used by each to make the determination, and their sensitivities (SE) and specificities (SP). SEs and SPs for the data in this table are calculated under the assumption that a query protein is a real allergen if its sequence matches a sequence in AllergenOnline v12 with 95% identity (see Section 2).

Assessment name	Potential allergenicity criteria	SE	SP
Established Method 1	Any alignment between any 80-amino-acid-long window on the query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 OR any sequence of six or more amino acids found in both the query protein and an allergen	0.992	0.147
Established Method 2	Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 OR any sequence of eight or more amino acids found in both the query protein and an allergen	0.951	0.898

Table 2

The 17 categories of potential criteria used to assess allergenicity (and the criteria used by each to make the determination). For each category, the specific condition(s) used to produce the highest specificity (SP) with sensitivity equal to (or greater than) that of Established Method 2 (0.951) is provided (or “none” if there are no specific conditions that can produce an equivalent sensitivity). Note that the sensitivity (SE) of Established Method 2 is 0.898. SEs and SPs for Established Method 2 and for the data in this table are calculated under the assumption that a query protein is a real allergen if it matches a sequence in AllergenOnline v12 with 95% identity (see Section 2).

Assessment name	Potential allergenicity criteria	Specific conditions used	SP
Potential Criteria Category 1	Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 (1 criterion tested)	None	None
Potential Criteria Category 2	Any sequence, at or above a specified length, found in both a query protein and an allergen (10 criteria tested)	None	None
Potential Criteria Category 3	Any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold (6 criteria tested, Table S1)	<i>E</i> -value = 10^{-6}	0.918
Potential Criteria Category 4	Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 OR any sequence, at or above a specified length, found in both a query protein and an allergen (10 criteria tested, Table S2)	Length = 13 aa	0.935
Potential Criteria Category 5	Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 AND any sequence, at or above a specified length, found in both a query protein and an allergen (10 criteria tested)	None	None
Potential Criteria Category 6	Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 OR any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold (10 criteria tested, Table S3)	<i>E</i> -value = 10^{-18}	0.929
Potential Criteria Category 7	Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 AND any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold (10 criteria tested)	None	None
Potential Criteria Category 8	Any sequence, at or above a specified length, found in both a query protein and an allergen OR any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold (60 criteria tested, Table S4)	Length = 9 aa <i>E</i> -value = 10^{-8}	0.919
Potential Criteria Category 9	Any sequence, at or above a specified length, found in both a query protein and an allergen AND any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold (60 criteria tested, Table S5)	Length = 6 aa <i>E</i> -value = 10^{-4}	0.910
Potential Criteria Category 10	Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 OR any sequence, at or above a specified length, found in both a query protein and an allergen OR any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold (60 criteria tested, Table S6)	Length = 13 aa <i>E</i> -value = 10^{-20}	0.931
Potential Criteria Category 11	[Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 OR any sequence, at or above a specified length, found in both a query protein and an allergen AND any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold (60 criteria tested, Table S7)]	Length = 13 aa <i>E</i> -value = 10^{-4}	0.942
Potential Criteria Category 12	[Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 OR any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold] AND any sequence, at or above a specified length, found in both a query protein and an allergen (60 criteria tested, Table S8)	Length = 6 aa <i>E</i> -value = 10^{-12}	0.926
Potential Criteria Category 13	[Any sequence, at or above a specified length, found in both a query protein and an allergen OR any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold] AND any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 (60 criteria tested)	None	None
Potential Criteria Category 14	[Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 AND any sequence, at or above a specified length, found in both a query protein and an allergen] OR any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold (60 criteria tested, Table S9)	Length = 8 aa <i>E</i> -value = 10^{-8}	0.925
Potential Criteria Category 15	[Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 AND any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold] OR any sequence, at or above a specified length, found in both a query protein and an allergen (60 criteria tested, Table S10)	Length = 13 aa <i>E</i> -value = 10^{-4}	0.942
Potential Criteria Category 16	[Any sequence, at or above a specified length, found in both a query protein and an allergen AND any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold] OR any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 (60 criteria tested, Table S11)	Length = 7 aa <i>E</i> -value = 10^{-18}	0.932
Potential Criteria Category 17	Any alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 AND any sequence, at or above a specified length, found in both a query protein and an allergen AND any alignment between a query protein and an allergen with an <i>E</i> -value at or below some specified threshold (60 criteria tested)	None	None

acids found in both the query protein and an allergen (Table 1, Established Method 1). The UniProtKB/Swiss-Prot plant proteins analyzed according to Established Method 1 produced a SE of 0.992 and a SP of 0.147, indicating an excellent ability to detect true positives but very poor ability to detect true negatives.

3.3. Allergenic assessment using Established Method 2 substantially reduces false positives

Because the original FAO/WHO guidelines produce such an abundance of false positives their implementation often proves

untenable. Alternate criteria designate a query protein potentially allergenic if there is (1) any alignment between the query protein and an allergen with $\geq 35\%$ sequence identity over an alignment length ≥ 80 or (2) some sequence of at least eight-amino-acids found in both the query protein and an allergen (Table 1, Established Method 2). In practice Established Method 2 is considerably more robust than Established Method 1, producing many fewer false positive results.

The UniProtKB/Swiss-Prot plant proteins were evaluated per Established Method 2, producing a SE of 0.951 and a SP of 0.898. The substantial increase in SP from the level resulting from the

use of Established Method 1 confirms that the quantity of false positives is dramatically reduced, accounting now for about 10% of the non-allergens. This reduction in false positives is achieved alongside a much smaller reduction in the ability to detect false negatives.

3.4. The predictive ability of Established Method 2 is determined by its “alignment” component rather than its “8-mer” component

Established Method 2 contains two components: A query protein is potentially allergenic if it either (1) aligns to an allergen with $\geq 35\%$ sequence identity over an alignment length ≥ 80 (the “alignment” component) or (2) contains some sequence of eight or more amino acids found in both it and an allergen (the “8-mer” component). These two components can be used separately as criteria to determine SE and SP (Table 2, Proposed Criteria Categories 1 and 2).

Analysis of the UniProtKB/Swiss-Prot protein set using only the alignment component produces a SE of 0.947 and a SP of 0.936. Consequently, the alignment component when used by itself, compared to using all of Established Method 2, significantly raises the SP (i.e., lowers the number of false positives) but also very slightly lowers the SE (i.e., raises the number of false negatives).

Assessment of allergenicity using only the 8-mer component produces a SE of 0.920 and an SP of 0.922. These numbers indicate a reduction in false positives compared to the level achieved when both the alignment and 8-mer components are employed simultaneously, but this reduction arises alongside a significant and unacceptable increase in false negatives.

A close examination of the data shows that precisely one of the 263 real allergens in the UniProtKB/Swiss-Prot query protein dataset is missed by the alignment component but found by the 8-mer component. This protein is CUCM1_CUCME, cucumisin, a serine protease from muskmelon. While it produces no alignments with $\geq 35\%$ sequence identity over an alignment length ≥ 80 to any allergens in its reduced AllergenOnline v12 database (see Section 2.4), it does possess a common 8-mer with four fungal allergens (gi numbers 289172, 21069093, 4587983, and 6684758). The identifying 8-mer contains one portion of the catalytic triad for these proteins (Arnorsdottir et al., 2005; Betzel et al., 1993; Chou et al., 2002; Chow et al., 2000; Frederick et al., 1993; Su et al., 1999). Therefore, even though there are no homologous sequences to this query protein in its reduced allergen database, a conserved feature in distantly-related allergens led to a correct hit. However, and importantly, the 8-mer that implicates the query protein is not its epitope; it is merely fortuitous that other serine proteases with a conserved catalytic motif happen to be allergens. Since the positive assessment is unrelated to the proteins' allergenic potentials, the presence of the match does not provide any evidence of the 8-mer component's general utility as a predictor of possible allergenicity. And since all of the other TPs resulting from the use of Established Method 2 were identified by the alignment component, it appears that the 8-mer component provides no predictive value at all.

3.5. A sequence of amino acids found in both a query protein and an allergen, of any examined length, does not predict allergenicity of the protein as well as Established Method 2

Length thresholds in the range from six to fifteen were used to assess SE and SP in the query protein datasets (Table 2, Proposed Criteria Category 2). Used as allergenic assessment criteria, no length threshold performed as well as Established Method 2. When the minimum length was set to six, SE rose to 0.989 while SP dropped to 0.150, indicating that almost all of the query proteins, whether allergens or not, contained a sequence of six or more

amino acids also found in an allergen. (Note that these values are almost the same as those obtained from the application of Established Method 1 above, suggesting that the high SE and very low SP obtained using Established Method 1 result principally from the presence of six-amino-acid-long (or longer) sequences found in both the query proteins and the allergens and not because 80-amino-acid-long “windows” derived from the query protein sequences produce numerous above-threshold alignments to the allergen sequences.) A length threshold of seven gave inferior performance relative to Established Method 2 for both SE and SP (data not shown). All other length thresholds, when compared to Established Method 2, gave inferior performance for SE, suggesting that these longer sequences found in both the query proteins and the allergens, by themselves, cannot adequately identify a potential allergen (data not shown).

3.6. E-value predicts protein allergenicity better than Established Method 2

One of the premier metrics quantifying the significance of a sequence alignment made by the FASTA alignment tool (Pearson and Lipman, 1988) is the *E*-value, which is indicative of the possibility that a particular sequence alignment could occur by chance. Lower *E*-values signify more biologically relevant alignments.

It has been suggested that an alignment between a query protein and an allergen possessing an *E*-value below $\sim 10^{-6}$ may indicate that the query protein is allergenic (Silvanovich et al., 2009). The proteins in the query protein dataset were assessed for allergenicity, and SE and SP determined, using *E*-value thresholds of 10^{-12} , 10^{-10} , 10^{-8} , 10^{-6} , 10^{-4} , or 10^{-2} as the sole criterion (Table 2, Proposed Criteria Category 3). At a threshold of 10^{-8} , SE is lower and SP is higher than that achieved with Established Method 2 (Table S1). At a threshold of 10^{-2} , SE is higher and SP is lower than that achieved with Established Method 2. Between these levels, both SE and SP are comparable to or higher than the values achieved with Established Method 2. In fact, at an *E*-value threshold of 10^{-6} , without any reduction in SE, SP rises to 0.918 (from 0.898 when Established Method 2 is used), equivalent to a decrease in false positives of 19.9%.

3.7. Some double-component criteria can be used to assess allergenicity with greater SE and SP than any single-component criterion

As shown above, some *E*-value thresholds, by themselves, can be used to assess allergenicity with comparable or greater SE and SP than Established Method 2 (which is a double-component criterion). Nevertheless, it was hoped that some other double-component criterion, would enhance performance even more. Thus, various double-component criteria were devised and tested for their ability to assess query proteins for allergenicity. Each double-component criterion consisted of two of the following three criteria: (1) a check for the presence of an alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 ; (2) a check for the presence of a sequence, at or above a specified length, found in both a query protein and an allergen; and (3) a check for the presence of an alignment between a query protein and an allergen with an *E*-value at or below some specified threshold.

In the first dual combination, a query protein is considered potentially allergenic if there is (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length of ≥ 80 or (2) some sequence above a certain length found in both the query protein and an allergen (Table 2, Proposed Criteria Category 4). (Note that when the length threshold is set to eight, this proposed criterion is the same as Established Method 2.) The analyses

show that this potential criteria category's best performance is achieved when the length threshold is raised from eight to thirteen – SP increases from 0.898 to 0.935, representing a 36.4% reduction in false positives, without any reduction in SE (Table S2).

The second dual combination designates a query protein potentially allergenic if there is (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length of ≥ 80 and (2) some sequence above a certain length found in both the query protein and an allergen (Table 2, Proposed Criteria Category 5). For all length thresholds explored (i.e., 6–15), SE deteriorated when compared to the results obtained with Established Method 2 (data not shown).

In the third dual combination, a protein is considered potentially allergenic if there is (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length of ≥ 80 or (2) an alignment between it and an allergen with an *E*-value below a specified limit (Table 2, Proposed Criteria Category 6). The best-performing *E*-value threshold in this proposed criteria category was 10^{-18} , producing a SP of 0.929, or a 30.8% reduction in false positives, with no reduction in SE, when compared to the values obtained by using Established Method 2 (Table S3).

The fourth dual combination calls a query protein a potential allergen if there is (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length of ≥ 80 and (2) an alignment between it and an allergen with an *E*-value below a specified limit (Table 2, Proposed Criteria Category 7). The results indicate that SE deteriorated, compared to the values obtained by using Established Method 2, across the full range (10^{-20} – 10^{-2}) of *E*-values explored (data not shown).

In the fifth dual combination, a protein is considered potentially allergenic if there is (1) some sequence above a certain length found in both it and an allergen or (2) an alignment between it and an allergen with an *E*-value below a specified limit (Table 2, Proposed Criteria Category 8). These analyses showed once again that some criteria employing double components can yield SPs exceeding those derived from the use of Established Method 2 while delivering no reduction in SE (Table S4). The best length and *E*-value thresholds producing this result were 9 and 10^{-8} , respectively: SP rose to 0.919, signifying a 20.5% reduction in false positives.

The final dual combination finds a query protein potentially allergenic when there is (1) some sequence above a certain length found in both it and an allergen and (2) an alignment between it and an allergen with an *E*-value below a specified limit (Table 2, Proposed Criteria Category 9). One combination of length and *E*-value thresholds in the ranges explored produced an increase in SP, relative to the level obtained with Established Method 2, without a reduction in SE (Table S5). At a length threshold of 6 and an *E*-value of 10^{-4} , SP rose to 0.910 – equivalent to a decrease of 11.7% in false positives.

3.8. Some triple-component criteria can be used to assess allergenicity with greater SE and SP than any single- or double-component criterion

Various double-component criteria, then, can be used to assess query proteins for allergenicity with performance that is superior to that obtained from any single-component criterion. This success suggested that triple-component criteria should also be investigated. All triple-component criteria examined included all three of the following components in various combinations: (1) a check for the presence of an alignment between a query protein and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 ; (2) a check for the presence of a sequence, at or above a specified length, found in both a query protein and an allergen; and (3) a check for an alignment between a query protein and an allergen with an *E*-value at or below some specified threshold.

The first triple combination declares a query protein potentially allergenic if there is (1) an alignment between it and an allergen hav-

ing $\geq 35\%$ sequence identity over an alignment length of ≥ 80 or (2) some sequence above a certain length found in both it and an allergen or (3) an alignment between it and an allergen with an *E*-value below a specified limit (Table 2, Proposed Criteria Category 10). Most of the length and *E*-value thresholds tested gave higher SP without sacrifice in SE when compared to the Established Method 2 results (Table S6). The best SP obtained without reduction in SP was 0.931, representing a 32.2% reduction in false positives, using a length threshold of 13 and an *E*-value threshold of 10^{-20} .

The second triple combination identifies a query protein as a potential allergen when there is either (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length of ≥ 80 or (2) some sequence above a certain length found in both it and an allergen, and (3) there is an alignment between it and an allergen with an *E*-value below a specified limit (Table 2, Proposed Criteria Category 11). Application of this criterion produced superior SP without reduction in SE, relative to Established Method 2, at 16 length and *E*-value thresholds (Table S7). The best thresholds, at 13 and 10^{-4} , respectively, gave a SP of 0.942, equal to a 43.4% reduction in false positives.

The use of this criterion for assessment of potential allergenicity produces the best results in this study, so it is useful here to show how some particular query proteins perform when evaluated under Established Method 2 and under this criterion. Of the 263 real allergens in the query protein dataset, the same 250 are identified correctly by both criteria (data not shown). However, as the statistics indicate, the number of FPs is much greater when Established Method 2 is used than when this proposed criterion is used (there are 2752 FPs under Established Method 2 and 1558 FPs under this criterion). Two randomly-selected query proteins that are FPs under Established Method 2 but TNs under this criterion are CYT4_ORYSJ and GAT22_ARATH. CYT4_ORYSJ is cysteine protease inhibitor 4 from *Oryza sativa* subsp. *japonica* (rice). It has an alignment to a phytocystatin from *Actinidia deliciosa* (kiwi) with 36.8% sequence identity over an alignment length of 106, but it has no alignments with an *E*-value less than 0.0012 and likewise no short sequences in common with any allergen. Under Established Method 2, the sequence identity and length of the alignment with the kiwi protein is sufficient to erroneously mark it a positive. However, under the proposed criterion considered here, because it must also have an alignment with an *E*-value less than or equal to 0.0001 to be considered a positive, it is properly labeled a negative.

GAT22_ARATH is (putatively) GATA transcription factor 22 from *Arabidopsis thaliana*. It has no alignments to any allergens with $\geq 35\%$ sequence identity over an alignment length of ≥ 80 nor with an *E*-value less than or equal to 0.0001. However it does contain a sequence of eight sequential glutamines (Qs) that is also found in numerous wheat allergens. Under Established Method 2, this string is sufficient to erroneously mark the query protein a positive, but because it has no alignment with sufficiently low *E*-value, the proposed criterion correctly calls it a negative.

The third triple combination finds a query protein potentially allergenic when there is either (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length of ≥ 80 or (2) an alignment between it and an allergen with an *E*-value below a specified limit, and (3) it contains some sequence above a certain length found in both it and an allergen (Table 2, Proposed Criteria Category 12). At a length threshold of six and at various *E*-value thresholds, these criteria in this category, compared to Established Method 2, produced superior SP without a reduction in SE, with the best improvement in SP occurring at an *E*-value threshold of 10^{-12} : SP at these thresholds was 0.926, representing a drop in false positives of 27.1% (Table S8).

In the fourth triple combination, a query protein is identified as a potential allergen if there is either (1) some sequence above a certain length found in both its sequence and an allergen sequence

or (2) an alignment between it and an allergen with an *E*-value below a specified limit, and (3) there is an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length of ≥ 80 (Table 2, Proposed Criteria Category 13). No proposed criteria in this category produced an SE comparable to the SE produced using Established Method 2 (data not shown).

The fifth triple combination calls a query protein potentially allergenic when there is both (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 and (2) some sequence above a certain length found in both it and an allergen, or (3) there is an alignment between it and an allergen with an *E*-value below a specified limit (Table 2, Proposed Criteria Category 14). Many of the proposed criteria in this category produced superior SP without a reduction in SE, compared to Established Method 2 (Table S9). The best length and *E*-value thresholds found (at 8 and 10^{-8} , respectively) gave a SP of 0.925, equal to a 26.6% reduction in false positives.

The sixth triple combination – a query protein is potentially allergenic when there is both (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 and (2) an alignment between it and an allergen with an *E*-value below a specified limit, or (3) it contains some sequence above a certain length found in both it and an allergen (Table 2, Proposed Criteria Category 15) – found a dozen cases where the length and *E*-value threshold used produced, compared to Established Method 2, superior SP without reduced SE (Table S10). The best instance of this occurred at a length threshold of 13 and an *E*-value threshold of 10^{-4} , yielding a SP of 0.942 or a reduction in false positives of 43.2%.

The seventh triple combination declares a query protein a potential allergen if there is both (1) some sequence above a certain length found in both it and an allergen and (2) an alignment between it and an allergen with an *E*-value below a specified limit, or (3) there is an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 (Table 2, Proposed Criteria Category 16). Eight of the length and *E*-value thresholds examined in this category, compared to Established Method 2, gave a superior SP with comparable or greater SE (Table S11). The best length and *E*-value thresholds (7 and 10^{-18} , respectively) produces a SP of 0.932, equal to a 33.2% reduction in false positives.

The last triple combination calls a query protein potentially allergenic when there is (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 and (2) some sequence of amino acids above a certain length found in both it and an allergen and (3) an alignment between it and an allergen with an *E*-value below a specified limit (Table 2, Proposed Criteria Category 17). Using these criteria, all of the length and *E*-value thresholds in the range explored gave reduced SE compared to the values obtained under Established Method 2 (data not shown).

3.9. Determination of real allergenicity at the 90% and 100% identity levels produces comparable or superior performance to the above data

All of the above tests were subsequently repeated with both a more-stringent (100% identity requirement) and less-stringent (90% identity requirement) test for actual allergenicity. For both the 100% and the 90% identity requirements, and for each of the 17 proposed criteria categories, the results were comparable to or superior to those obtained at the 95% level (data not shown). Therefore, the data presented above produces the most conservative interpretation available from analyses at these three levels.

4. Discussion

This study determined the sensitivities and specificities resulting from the use of two established methods and hundreds of possible methods (in 17 categories) to assess transgenic proteins, *in*

silico, for their potential allergenicity. The earliest guidelines (FAO/WHO, 2001) (Established Method 1) unfortunately flagged an enormous fraction of non-allergens as possible problem proteins: 85% of the non-allergens in the query protein datasets evaluated here were found to be potentially allergenic by this method. This finding is in accord with previous research by Lim et al. (2008) demonstrating that Established Method 1, while correctly identifying most of the true positives, nevertheless also produced a wealth of false positives. The use of Established Method 2 as a test for potential allergenicity substantially reduces the false positive rate, from $\sim 85\%$ to $\sim 10\%$, while only slightly lowering the ability to detect true positives. The data here show how the assessment for the detection of possible allergenicity can be modified in ways that further reduce the fraction of false positives while maintaining Established Method 2's ability to detect true positives. This can be accomplished using numerous criteria as shown in Section 3. However, the most effective criterion found declares a query protein potentially allergenic if there is either (1) an alignment between it and an allergen having $\geq 35\%$ sequence identity over an alignment length ≥ 80 or (2) some sequence of 13 or more amino acids found both in it and an allergen, and (3) there is an alignment between it and an allergen with an *E*-value $\leq 10^{-4}$. When both this proposed criterion and Established Method 2 were applied against the same query protein dataset, the proposed criterion produced fewer than 6% false positives – a reduction of $\sim 43\%$ from Established Method 2's 10% false positive level – while causing no increase in false negatives. This reduction is highly significant, as positive hits often require extensive experimental follow-up to determine the protein's actual risk – which can be both expensive and time-consuming. A change from current practice to this proposed method could therefore help reduce developmental costs and time-to-market for transgenic crops and has no obvious downside.

Any proteins in the query protein dataset that do not match an allergen in the allergen database at one of the three identity threshold levels (depending on the test) are assumed to be non-allergens. And yet, it could be that the allergen database is incomplete and that in some future incarnation of the database, containing more allergens, there would be more query proteins designated real allergens and fewer designated non-allergens. How might this affect the results? Because SE is the ratio of TPs to the total number of real allergens and SP is the ratio of TNs to the total number of non-allergens, our SE and SP values for each criterion would not change provided the fraction of newly designated real allergens that assessed positively under that criterion remained constant (i.e., the newly-designated real allergens behaved the same as the previously-assessed allergens in the way they distribute as TPs, FNs, TNs, and FPs). This is not guaranteed, but one way to test this “in reverse” is to repeat all of our tests with a prior version of the database. We completed an identical set of evaluations using the previous version of the database (AllergenOnline v11), which contains 1491 sequences rather than the 1603 sequences in AllergenOnline v12, and obtained results that are very similar to those provided in this paper (data not shown).

There are surely many additional allergenic assessment criteria that could be explored. For example, the alignment identity threshold was kept constant in this study at 35%, though a variety of thresholds might have been examined – alone and in various combinations with contiguous identity length and *E*-value thresholds. A second possible permutation on the assessment criteria involves the location of the short sequence found in both the query protein and an allergen. In Established Method 2, and in all of the data described in results and shown in the tables, it is sufficient for such a short sequence to exist irrespective of its location (i.e., it does not need to be located within an alignment between the query protein and an allergen meeting or exceeding the 35% sequence identity/80

length-of-alignment requirement). However, it would seem to be a reasonable hypothesis that the short sequence represents an epitope that might be expected to appear in the context of a longer region of similarity, so that restricting its location to the broader alignment might improve the SE above the levels found. Third, if the short sequence actually does represent an epitope in broader context, then both it and the longer alignment would be expected to match the same allergen – and a criterion could be developed that tests for allergenic potential with this restriction as a requirement. Fourth, an allergen is typically composed of both allergenic and non-allergenic domains. Shrimp tropomyosin, for example, has five allergenic domains separated by intervening sequences (Ayuso et al., 2002; Reese et al., 2005). If an above-threshold alignment between a query protein and an allergen or an above-length-threshold sequence found in both the query protein and an allergen arises within the portion of an allergen that does not contain an allergenic domain, that alignment probably should not be used to trigger a positive potential allergenicity assessment.

All of these modifications would be interesting to test. Nevertheless, the present report shows that there are numerous assessment criteria – some only slight variations on Established Method 2 (e.g., simply raising the length threshold for a sequence found in both the query protein and an allergen from eight to thirteen) – that will maintain the current ability to detect possible allergenic risk while also reducing the false positive rate and allowing efficacious crop traits to be brought to market in a more expedient manner. The potential criteria proposed here should be tested by other researchers, and if similar results are achieved they should be strongly considered for incorporation into future guidelines for the evaluation of GM crop safety.

Conflict of interest

All authors are employed by DuPont Pioneer.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.yrtph.2013.08.001>.

References

- Aalberse, R.C., 2000. Structural biology of allergens. *J. Allergy Clin. Immunol.* 106, 228–238.
- Aalberse, R.C., Akkerdaas, J., van Ree, R., 2001. Cross-reactivity of IgE antibodies to allergens. *Allergy* 56, 478–490.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Arnorsdottir, J., Kristjansson, M.M., Ficner, R., 2005. Crystal structure of a subtilisin-like serine proteinase from a psychrotrophic *Vibrio* species reveals structural aspects of cold adaptation. *FEBS J.* 272, 832–845.
- Ayuso, R., Lehrer, S.B., Reese, G., 2002. Identification of continuous, allergenic regions of the major shrimp allergen Pen a 1 (tropomyosin). *Int. Arch. Allergy Immunol.* 127, 27–37.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16, 412–424.
- Betzel, C., Singh, T.P., Visanji, M., Peters, K., Fittkau, S., Saenger, W., Wilson, K.S., 1993. Structure of the complex of proteinase K with a substrate analogue hexapeptide inhibitor at 2.2-Å resolution. *J. Biol. Chem.* 268, 15854–15858.
- Burset, M., Guigo, R., 1996. Evaluation of gene structure prediction programs. *Genomics* 34, 353–367.
- Chou, H., Lai, H.Y., Tam, M.F., Chou, M.Y., Wang, S.R., Han, S.H., Shen, H.D., 2002. CDNA cloning, biological and immunological characterization of the alkaline serine protease major allergen from *Penicillium chrysogenum*. *Int. Arch. Allergy Immunol.* 127, 15–26.
- Chow, L.P., Chiou, S.H., Hsiao, M.C., Yu, C.J., Chiang, B.L., 2000. Characterization of Pen n 13, a major allergen from the mold *Penicillium notatum*. *Biochem. Biophys. Res. Commun.* 269, 14–20.
- Codex Alimentarius Commission, 2009. Foods Derived from Modern Biotechnology, second ed. Food and Agriculture Organization of the United Nations, World Health Organization, Rome.
- Cressman, R.F., Ladics, G., 2009. Further evaluation of the utility of “sliding window” FASTA in predicting cross-reactivity with allergenic proteins. *Regul. Toxicol. Pharmacol.* 54, S20–S25.
- EFSA, 2011. Guidance for risk assessment of food and feed from genetically modified plants. *EFSA J.* 9, 2150.
- FAO/WHO, 2001. Evaluation of allergenicity of genetically modified foods: report of a joint FAO/WHO expert consultation on allergenicity of foods derived from biotechnology. Food and Agriculture Organization of the United Nations, Rome, January 22–25, 2001.
- Frederick, G.D., Rombouts, P., Buxton, F.P., 1993. Cloning and characterisation of pepC, a gene encoding a serine protease from *Aspergillus niger*. *Gene* 125, 57–64.
- Goodman, R.E., 2008. Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. *Food Chem. Toxicol.* 46, S24–S34.
- Goodman, R.E., Tetteh, A.O., 2011. Suggested improvements for the allergenicity assessment of genetically modified plants used in foods. *Curr. Allergy Asthma Rep.* 11, 317–324.
- Goodman, R.E., Vieths, S., Sampson, H.A., Hill, D., Ebisawa, M., Taylor, S.L., van Ree, R., 2008. Allergenicity assessment of genetically modified crops – what makes sense? *Nat. Biotechnol.* 26, 73–81.
- Harper, B., McClain, S., Ganko, E.W., 2012. Interpreting the biological relevance of bioinformatic analyses with T-DNA sequence for protein allergenicity. *Regul. Toxicol. Pharmacol.* 63, 426–432.
- Herman, R.A., Song, P., Thirumalaiswamysekhar, A., 2009. Value of eight-amino-acid matches in predicting the allergenicity status of proteins: an empirical bioinformatic investigation. *Clin. Mol. Allergy* 7, 9.
- Klinglmayr, E., Hauser, M., Zimmermann, F., Dissertori, O., Lackner, P., Wopfner, N., Ferreira, F., Wallner, M., 2009. Identification of B-cell epitopes of Bet v 1 involved in cross-reactivity with food allergens. *Allergy* 64, 647–651.
- Ladics, G.S., Bannon, G.A., Silvanovich, A., Cressman, R.F., 2007. Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens. *Mol. Nutr. Food Res.* 51, 985–998.
- Ladics, G.S., Cressman, R.F., Herouet-Guicheney, C., Herman, R.A., Privalle, L., Song, P., Ward, J.M., McClain, S., 2011. Bioinformatics and the allergy assessment of agricultural biotechnology products: industry practices and recommendations. *Regul. Toxicol. Pharmacol.* 60, 46–53.
- Lim, S.J., Tong, J.C., Chew, F.T., Tammi, M.T., 2008. The value of position-specific scoring matrices for assessment of protein allergenicity. *BMC Bioinform.* 9 (Suppl. 12), S21.
- Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132, 185–219.
- Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.
- Reese, G. et al., 2005. Reduced allergenic potency of VR9-1, a mutant of the major shrimp allergen Pen a 1 (tropomyosin). *J. Immunol.* 175, 8354–8364.
- Silvanovich, A., Bannon, G., McClain, S., 2009. The use of E-scores to determine the quality of protein alignments. *Regul. Toxicol. Pharmacol.* 54, S26–S31.
- Su, N.Y., Yu, C.J., Shen, H.D., Pan, F.M., Chow, L.P., 1999. Pen c 1, a novel enzymic allergen protein from *Penicillium citrinum*. Purification, characterization, cloning and expression. *Eur. J. Biochem. FEBS* 261, 115–123.
- Young, G.J., Zhang, S., Mirsky, H.P., Cressman, R.F., Cong, B., Ladics, G.S., Zhong, C.X., 2012. Assessment of possible allergenicity of hypothetical ORFs in common food crops using current bioinformatic guidelines and its implications for the safety assessment of GM crops. *Food Chem. Toxicol.*

Table S1. SEs and SPs for Potential Criteria Category 3 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has an *E*-value at or below some specified value) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 3 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 3 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 3 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value threshold.

E-value					
1E-12	1E-10	1E-8	1E-6	1E-4	1E-2
SE = 0.928 SP = 0.942	SE = 0.935 SP = 0.936	SE = 0.943 SP = 0.929	SE = 0.951 SP = 0.918	SE = 0.962 SP = 0.906	SE = 0.977 SP = 0.874

Table S2. SEs and SPs for Potential Criteria Category 4 (i.e., a query protein is declared potentially allergenic if (any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater) OR (some sequence of amino acids of some minimum specified length exists in both the query protein and an allergen)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 4 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 4 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 4 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing length threshold.

"Word" Length									
6	7	8	9	10	11	12	13	14	15
SE = 0.992 SP = 0.149	SE = 0.958 SP = 0.725	SE = 0.951 SP = 0.898	SE = 0.951 SP = 0.926	SE = 0.951 SP = 0.932	SE = 0.951 SP = 0.934	SE = 0.951 SP = 0.935	SE = 0.951 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935

Table S3. SEs and SPs for Potential Criteria Category 6 (i.e., a query protein is declared potentially allergenic if (any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater) OR (any alignment to an allergen has an *E*-value at or below some specified value)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 6 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 6 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 6 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value threshold.

E-value									
1E-20	1E-18	1E-16	1E-14	1E-12	1E-10	1E-8	1E-6	1E-4	1E-2
SE = 0.947 SP = 0.931	SE = 0.951 SP = 0.929	SE = 0.951 SP = 0.928	SE = 0.951 SP = 0.926	SE = 0.954 SP = 0.924	SE = 0.958 SP = 0.921	SE = 0.958 SP = 0.916	SE = 0.958 SP = 0.909	SE = 0.962 SP = 0.899	SE = 0.977 SP = 0.870

Table S4. SEs and SPs for Potential Criteria Category 8 (i.e., a query protein is declared potentially allergenic if (some sequence of amino acids of some minimum specified length exists in both the query protein and an allergen) OR (any alignment to an allergen has an *E*-value at or below some specified value)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 8 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 8 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 8 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value and length thresholds.

		"Word" Length									
		6	7	8	9	10	11	12	13	14	15
E-value	1E-12	SE = 0.989 SP = 0.149	SE = 0.954 SP = 0.727	SE = 0.939 SP = 0.902	SE = 0.935 SP = 0.931	SE = 0.932 SP = 0.937	SE = 0.932 SP = 0.940	SE = 0.932 SP = 0.941	SE = 0.932 SP = 0.941	SE = 0.928 SP = 0.941	SE = 0.928 SP = 0.942
	1E-10	SE = 0.989 SP = 0.149	SE = 0.962 SP = 0.724	SE = 0.947 SP = 0.898	SE = 0.943 SP = 0.926	SE = 0.939 SP = 0.932	SE = 0.939 SP = 0.934	SE = 0.939 SP = 0.935	SE = 0.939 SP = 0.935	SE = 0.935 SP = 0.936	SE = 0.935 SP = 0.936
	1E-8	SE = 0.992 SP = 0.148	SE = 0.966 SP = 0.721	SE = 0.954 SP = 0.892	SE = 0.951 SP = 0.919	SE = 0.947 SP = 0.925	SE = 0.947 SP = 0.927	SE = 0.947 SP = 0.928	SE = 0.947 SP = 0.928	SE = 0.943 SP = 0.928	SE = 0.943 SP = 0.928
	1E-6	SE = 0.992 SP = 0.148	SE = 0.966 SP = 0.716	SE = 0.958 SP = 0.884	SE = 0.954 SP = 0.910	SE = 0.954 SP = 0.915	SE = 0.954 SP = 0.917	SE = 0.954 SP = 0.918	SE = 0.954 SP = 0.918	SE = 0.951 SP = 0.918	SE = 0.951 SP = 0.918
	1E-4	SE = 0.992 SP = 0.146	SE = 0.966 SP = 0.708	SE = 0.962 SP = 0.873	SE = 0.962 SP = 0.898	SE = 0.962 SP = 0.903	SE = 0.962 SP = 0.905	SE = 0.962 SP = 0.906	SE = 0.962 SP = 0.906	SE = 0.962 SP = 0.906	SE = 0.962 SP = 0.906
	1E-2	SE = 0.992 SP = 0.143	SE = 0.981 SP = 0.689	SE = 0.977 SP = 0.845	SE = 0.977 SP = 0.868	SE = 0.977 SP = 0.872	SE = 0.977 SP = 0.873	SE = 0.977 SP = 0.874	SE = 0.977 SP = 0.874	SE = 0.977 SP = 0.874	SE = 0.977 SP = 0.874

Table S5. SEs and SPs for Potential Criteria Category 9 (i.e., a query protein is declared potentially allergenic if (some sequence of amino acids of some minimum specified length exists in both the query protein and an allergen) AND (any alignment to an allergen has an *E*-value at or below some specified value)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 9 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 9 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 9 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value and length thresholds.

		"Word" Length									
		6	7	8	9	10	11	12	13	14	15
E-value	1E-12	SE = 0.928 SP = 0.943	SE = 0.916 SP = 0.955	SE = 0.909 SP = 0.962	SE = 0.890 SP = 0.968	SE = 0.882 SP = 0.971	SE = 0.867 SP = 0.974	SE = 0.859 SP = 0.978	SE = 0.852 SP = 0.981	SE = 0.852 SP = 0.983	SE = 0.837 SP = 0.986
	1E-10	SE = 0.935 SP = 0.938	SE = 0.916 SP = 0.952	SE = 0.909 SP = 0.960	SE = 0.890 SP = 0.967	SE = 0.882 SP = 0.971	SE = 0.867 SP = 0.974	SE = 0.859 SP = 0.978	SE = 0.852 SP = 0.981	SE = 0.852 SP = 0.983	SE = 0.837 SP = 0.986
	1E-8	SE = 0.939 SP = 0.931	SE = 0.920 SP = 0.948	SE = 0.909 SP = 0.958	SE = 0.890 SP = 0.966	SE = 0.882 SP = 0.970	SE = 0.867 SP = 0.973	SE = 0.859 SP = 0.978	SE = 0.852 SP = 0.981	SE = 0.852 SP = 0.983	SE = 0.837 SP = 0.985
	1E-6	SE = 0.947 SP = 0.921	SE = 0.928 SP = 0.942	SE = 0.913 SP = 0.956	SE = 0.894 SP = 0.965	SE = 0.882 SP = 0.969	SE = 0.867 SP = 0.972	SE = 0.859 SP = 0.977	SE = 0.852 SP = 0.980	SE = 0.852 SP = 0.983	SE = 0.837 SP = 0.985
	1E-4	SE = 0.958 SP = 0.910	SE = 0.939 SP = 0.938	SE = 0.920 SP = 0.955	SE = 0.897 SP = 0.964	SE = 0.886 SP = 0.969	SE = 0.871 SP = 0.972	SE = 0.863 SP = 0.977	SE = 0.856 SP = 0.980	SE = 0.852 SP = 0.983	SE = 0.837 SP = 0.985
	1E-2	SE = 0.973 SP = 0.881	SE = 0.939 SP = 0.926	SE = 0.920 SP = 0.952	SE = 0.897 SP = 0.963	SE = 0.886 SP = 0.968	SE = 0.871 SP = 0.972	SE = 0.863 SP = 0.977	SE = 0.856 SP = 0.980	SE = 0.852 SP = 0.982	SE = 0.837 SP = 0.985

Table S6. SEs and SPs for Potential Criteria Category 10 (i.e., a query protein is declared potentially allergenic if (any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater) OR (some sequence of amino acids of some minimum specified length exists in both the query protein and an allergen) OR (any alignment to an allergen has an *E*-value at or below some specified value)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists between the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 10 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 10 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 10 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value and length thresholds.

		"Word" Length									
		6	7	8	9	10	11	12	13	14	15
E-value	1E-20	SE = 0.992 SP = 0.149	SE = 0.958 SP = 0.724	SE = 0.951 SP = 0.896	SE = 0.951 SP = 0.923	SE = 0.951 SP = 0.928	SE = 0.951 SP = 0.930	SE = 0.951 SP = 0.931	SE = 0.951 SP = 0.931	SE = 0.947 SP = 0.931	SE = 0.947 SP = 0.931
	1E-18	SE = 0.992 SP = 0.149	SE = 0.958 SP = 0.723	SE = 0.954 SP = 0.894	SE = 0.954 SP = 0.921	SE = 0.954 SP = 0.926	SE = 0.954 SP = 0.928	SE = 0.954 SP = 0.929	SE = 0.954 SP = 0.929	SE = 0.951 SP = 0.929	SE = 0.951 SP = 0.929
	1E-16	SE = 0.992 SP = 0.149	SE = 0.958 SP = 0.722	SE = 0.954 SP = 0.893	SE = 0.954 SP = 0.919	SE = 0.954 SP = 0.924	SE = 0.954 SP = 0.926	SE = 0.954 SP = 0.927	SE = 0.954 SP = 0.927	SE = 0.951 SP = 0.927	SE = 0.951 SP = 0.927
	1E-14	SE = 0.992 SP = 0.149	SE = 0.958 SP = 0.721	SE = 0.954 SP = 0.892	SE = 0.954 SP = 0.918	SE = 0.954 SP = 0.923	SE = 0.954 SP = 0.925	SE = 0.954 SP = 0.925	SE = 0.954 SP = 0.926	SE = 0.951 SP = 0.926	SE = 0.951 SP = 0.926
	1E-12	SE = 0.992 SP = 0.149	SE = 0.962 SP = 0.720	SE = 0.958 SP = 0.890	SE = 0.958 SP = 0.916	SE = 0.958 SP = 0.921	SE = 0.958 SP = 0.923	SE = 0.958 SP = 0.923	SE = 0.958 SP = 0.924	SE = 0.954 SP = 0.924	SE = 0.954 SP = 0.924
	1E-10	SE = 0.992 SP = 0.149	SE = 0.966 SP = 0.718	SE = 0.962 SP = 0.887	SE = 0.962 SP = 0.913	SE = 0.962 SP = 0.918	SE = 0.962 SP = 0.919	SE = 0.962 SP = 0.920	SE = 0.962 SP = 0.920	SE = 0.958 SP = 0.920	SE = 0.958 SP = 0.921

Table S7. SEs and SPs for Potential Criteria Category 11 (i.e., a query protein is declared potentially allergenic if [(any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater) OR (some sequence of amino acids of some minimum specified length exists in both the query protein and an allergen)] AND (any alignment to an allergen has an *E*-value at or below some specified value)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 11 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 11 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 11 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value and length thresholds.

		"Word" Length									
		6	7	8	9	10	11	12	13	14	15
E-value	1E-12	SE = 0.928 SP = 0.943	SE = 0.924 SP = 0.948	SE = 0.920 SP = 0.956	SE = 0.920 SP = 0.953	SE = 0.920 SP = 0.953	SE = 0.920 SP = 0.953	SE = 0.920 SP = 0.954	SE = 0.920 SP = 0.954	SE = 0.920 SP = 0.954	SE = 0.920 SP = 0.954
	1E-10	SE = 0.935 SP = 0.937	SE = 0.928 SP = 0.944	SE = 0.924 SP = 0.948	SE = 0.924 SP = 0.950	SE = 0.924 SP = 0.950	SE = 0.924 SP = 0.951	SE = 0.924 SP = 0.951	SE = 0.924 SP = 0.951	SE = 0.924 SP = 0.951	SE = 0.924 SP = 0.951
	1E-8	SE = 0.943 SP = 0.930	SE = 0.935 SP = 0.938	SE = 0.932 SP = 0.943	SE = 0.932 SP = 0.946	SE = 0.932 SP = 0.947	SE = 0.932 SP = 0.947	SE = 0.932 SP = 0.947	SE = 0.932 SP = 0.948	SE = 0.932 SP = 0.948	SE = 0.932 SP = 0.948
	1E-6	SE = 0.951 SP = 0.920	SE = 0.943 SP = 0.932	SE = 0.939 SP = 0.940	SE = 0.939 SP = 0.943	SE = 0.939 SP = 0.944	SE = 0.939 SP = 0.944	SE = 0.939 SP = 0.944	SE = 0.939 SP = 0.945	SE = 0.939 SP = 0.945	SE = 0.939 SP = 0.945
	1E-4	SE = 0.962 SP = 0.909	SE = 0.954 SP = 0.926	SE = 0.951 SP = 0.936	SE = 0.951 SP = 0.941	SE = 0.951 SP = 0.941	SE = 0.951 SP = 0.942	SE = 0.951 SP = 0.942	SE = 0.951 SP = 0.942	SE = 0.947 SP = 0.942	SE = 0.947 SP = 0.943
	1E-2	SE = 0.977 SP = 0.881	SE = 0.954 SP = 0.913	SE = 0.951 SP = 0.931	SE = 0.951 SP = 0.937	SE = 0.951 SP = 0.938	SE = 0.951 SP = 0.938	SE = 0.951 SP = 0.939	SE = 0.951 SP = 0.939	SE = 0.947 SP = 0.939	SE = 0.947 SP = 0.939

Table S8. SEs and SPs for Potential Criteria Category 12 (i.e., a query protein is declared potentially allergenic if [(any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater) OR (any alignment to an allergen has an *E*-value at or below some specified value)] AND (some sequence of amino acids of some minimum specified length exists in both the query protein and an allergen)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 12 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 12 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 12 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value and length thresholds.

		"Word" Length									
		6	7	8	9	10	11	12	13	14	15
E-value	1E-12	SE = 0.951 SP = 0.926	SE = 0.935 SP = 0.944	SE = 0.916 SP = 0.956	SE = 0.894 SP = 0.964	SE = 0.882 SP = 0.969	SE = 0.867 SP = 0.972	SE = 0.859 SP = 0.977	SE = 0.852 SP = 0.980	SE = 0.852 SP = 0.983	SE = 0.837 SP = 0.985
	1E-10	SE = 0.954 SP = 0.923	SE = 0.935 SP = 0.943	SE = 0.916 SP = 0.956	SE = 0.894 SP = 0.964	SE = 0.882 SP = 0.969	SE = 0.867 SP = 0.972	SE = 0.859 SP = 0.977	SE = 0.852 SP = 0.980	SE = 0.852 SP = 0.983	SE = 0.837 SP = 0.985
	1E-8	SE = 0.954 SP = 0.919	SE = 0.935 SP = 0.941	SE = 0.916 SP = 0.955	SE = 0.894 SP = 0.964	SE = 0.882 SP = 0.969	SE = 0.867 SP = 0.972	SE = 0.859 SP = 0.977	SE = 0.852 SP = 0.980	SE = 0.852 SP = 0.983	SE = 0.837 SP = 0.985
	1E-6	SE = 0.954 SP = 0.912	SE = 0.935 SP = 0.937	SE = 0.916 SP = 0.954	SE = 0.894 SP = 0.963	SE = 0.882 SP = 0.968	SE = 0.867 SP = 0.972	SE = 0.859 SP = 0.977	SE = 0.852 SP = 0.980	SE = 0.852 SP = 0.982	SE = 0.837 SP = 0.985
	1E-4	SE = 0.958 SP = 0.903	SE = 0.939 SP = 0.934	SE = 0.920 SP = 0.953	SE = 0.897 SP = 0.963	SE = 0.886 SP = 0.968	SE = 0.871 SP = 0.972	SE = 0.863 SP = 0.977	SE = 0.856 SP = 0.980	SE = 0.852 SP = 0.982	SE = 0.837 SP = 0.985
	1E-2	SE = 0.973 SP = 0.878	SE = 0.939 SP = 0.924	SE = 0.920 SP = 0.951	SE = 0.897 SP = 0.963	SE = 0.886 SP = 0.968	SE = 0.871 SP = 0.972	SE = 0.863 SP = 0.977	SE = 0.856 SP = 0.980	SE = 0.852 SP = 0.982	SE = 0.837 SP = 0.985

Table S9. SEs and SPs for Potential Criteria Category 14 (i.e., a query protein is declared potentially allergenic if [(any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater) AND (some sequence of amino acids of some minimum specified length exists in both the query protein and an allergen)] OR (any alignment to an allergen has an *E*-value at or below some specified value)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 14 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 14 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 14 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value and length thresholds.

[illegible]

Table S10. SEs and SPs for Potential Criteria Category 15 (i.e., a query protein is declared potentially allergenic if [(any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater) AND (any alignment to an allergen has an *E*-value at or below some specified value)] OR (some sequence of amino acids of some minimum specified length exists in both the query protein and an allergen)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 15 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 15 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 15 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value and length thresholds.

		"Word" Length									
		6	7	8	9	10	11	12	13	14	15
E-value	1E-12	SE = 0.989 SP = 0.150	SE = 0.951 SP = 0.733	SE = 0.932 SP = 0.911	SE = 0.928 SP = 0.941	SE = 0.924 SP = 0.948	SE = 0.924 SP = 0.951	SE = 0.924 SP = 0.952	SE = 0.924 SP = 0.953	SE = 0.920 SP = 0.953	SE = 0.920 SP = 0.953
	1E-10	SE = 0.989 SP = 0.150	SE = 0.954 SP = 0.732	SE = 0.935 SP = 0.909	SE = 0.932 SP = 0.939	SE = 0.928 SP = 0.946	SE = 0.928 SP = 0.948	SE = 0.928 SP = 0.949	SE = 0.928 SP = 0.950	SE = 0.924 SP = 0.950	SE = 0.924 SP = 0.951
	1E-8	SE = 0.992 SP = 0.150	SE = 0.958 SP = 0.730	SE = 0.943 SP = 0.907	SE = 0.939 SP = 0.937	SE = 0.935 SP = 0.943	SE = 0.935 SP = 0.945	SE = 0.935 SP = 0.946	SE = 0.935 SP = 0.947	SE = 0.932 SP = 0.947	SE = 0.932 SP = 0.947
	1E-6	SE = 0.992 SP = 0.149	SE = 0.958 SP = 0.730	SE = 0.947 SP = 0.905	SE = 0.943 SP = 0.935	SE = 0.943 SP = 0.941	SE = 0.943 SP = 0.943	SE = 0.943 SP = 0.944	SE = 0.943 SP = 0.944	SE = 0.939 SP = 0.944	SE = 0.939 SP = 0.945
	1E-4	SE = 0.992 SP = 0.149	SE = 0.958 SP = 0.728	SE = 0.951 SP = 0.903	SE = 0.951 SP = 0.933	SE = 0.951 SP = 0.938	SE = 0.951 SP = 0.941	SE = 0.951 SP = 0.941	SE = 0.951 SP = 0.942	SE = 0.947 SP = 0.942	SE = 0.947 SP = 0.942
	1E-2	SE = 0.992 SP = 0.149	SE = 0.958 SP = 0.727	SE = 0.951 SP = 0.901	SE = 0.951 SP = 0.930	SE = 0.951 SP = 0.935	SE = 0.951 SP = 0.938	SE = 0.951 SP = 0.939	SE = 0.951 SP = 0.939	SE = 0.947 SP = 0.939	SE = 0.947 SP = 0.939

Table S11. SEs and SPs for Potential Criteria Category 16 (i.e., a query protein is declared potentially allergenic if [(some sequence of amino acids of some minimum specified length exists in both the query protein and an allergen) AND (any alignment to an allergen has an *E*-value at or below some specified value)] OR (any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater)) to assess allergenicity of each protein in the query protein set. These SEs and SPs are compared to those arising from the use of Established Method 2 (i.e., a query protein is declared potentially allergenic if any alignment to an allergen has a length of 80 or greater amino acids and a sequence identity of 35% or greater OR some sequence of at least eight amino acids exists in both the query protein and an allergen) on the same proteins. Blue fill indicates that Potential Criterion 16 is inferior to Established Method 2 for SE but equal to or superior to Established Method 2 for SP. Green fill indicates that Potential Criterion 16 is equal to or superior to Established Method 2 for both SE and SP. Yellow fill indicates that Potential Criterion 16 is equal to or superior to Established Method 2 for SE but inferior to Established Method 2 for SP. The red square highlights the best-performing *E*-value and length thresholds.

		"Word" Length									
		6	7	8	9	10	11	12	13	14	15
E-value	1E-20	SE = 0.947 SP = 0.931	SE = 0.947 SP = 0.933	SE = 0.947 SP = 0.934	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936
	1E-18	SE = 0.951 SP = 0.930	SE = 0.951 SP = 0.932	SE = 0.947 SP = 0.933	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936
	1E-16	SE = 0.951 SP = 0.928	SE = 0.951 SP = 0.931	SE = 0.947 SP = 0.933	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936
	1E-14	SE = 0.951 SP = 0.926	SE = 0.951 SP = 0.930	SE = 0.947 SP = 0.933	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936
	1E-12	SE = 0.954 SP = 0.924	SE = 0.951 SP = 0.929	SE = 0.947 SP = 0.932	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936
	1E-10	SE = 0.958 SP = 0.921	SE = 0.951 SP = 0.928	SE = 0.947 SP = 0.932	SE = 0.947 SP = 0.934	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.935	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936	SE = 0.947 SP = 0.936